面向篇章机器翻译的英汉翻译单位和翻译模型研究 *

宋柔，葛诗利

广东外语外贸大学外语研究与语言服务协同创新中心，广东广州 510420

摘要：篇章机器翻译的首要问题是确定翻译单位。基于汉语和英语的语言知识和英汉翻译的实践，本文提出面向篇章机器翻译的基本单位和复合单位的双层单位体系，讨论了这两种单位支持篇章翻译应满足的性质，并据此勾画了篇章机器翻译的拆分、翻译、装配三步模型（PTA 模型）。本文提出，汉语篇章机器翻译的复合单位为广义话题结构对应的文本块，基本单位则是根据广义话题结构流水模型得到的话题自足句；英语篇章机器翻译的复合单位为句号句，基本单位为 naming-telling 小句（NT 小句），即指称性成分加上对它的陈述或后修饰成分所构成的小句。本文展示了在这样的翻译单位体系下采用 PTA 模型的英汉翻译过程实例，规划了面向篇章翻译的英汉小句对齐语料库的建设任务，讨论了 PTA 模型的可行性。

关键词：翻译单位；翻译模型；广义话题结构；naming-telling 小句

English-Chinese Translation Unit and Translation Model for Discourse-Based Machine Translation

Rou Song, Shili Ge

Guangdong Collaborative Innovation Center for Language Research & Services, Guangdong University of Foreign Studies, Guangzhou, Guangdong 510420, China

Abstract: The primary issue in discourse-based machine translation (MT) is to define translation unit. Based on English and Chinese linguistic knowledge and English-Chinese translation practice, we propose a double level system of translation unit, including basic unit and compound unit, for discourse-based MT. We further explore the required properties of these two types of units and construct a three-step model of parsing, translating and assembling (PTA model) in support of discourse-based MT. This paper indicates that the compound unit for Chinese discourse-based MT is the text in correspondence to generalized topic structure and the basic unit is the topic sufficient sentence derived from the stream model of generalized topic structure, while the compound unit for English is the traditional sentence and the basic unit is the naming-telling clause (NT clause), namely, the clause constructed with the referential component and its description or post-modification component. This paper exhibits the process of English-Chinese translation with an example under the theory of the double level translation unit system and PTA model. We set up the plan for the construction of English-Chinese clause aligned corpus for discourse-based MT and discuss the feasibility of PTA model.

Key words: Translation Unit; Translation Model; Generalized Topic Structure; Naming-Telling Clause

1. 引言

机器翻译是既有巨大需求又有巨大困难的自然语言处理课题，国内外已有多年的深入研究（冯志伟 2003，刘群 2012，宗成庆 2013）。基于规则的方法难以应对千变万化的语言现象，于是基于统计的方法应运而生。但句子层面的统计方法难以照顾远程的上下文相关关系，于是基于篇章的机器翻译成为当前的研究热点（张民 2014）。

篇章机器翻译的研究，目前多关注于利用篇章的词汇衔接信息和逻辑连贯信息，改进基于句子的机器翻译，包括词汇义项选择、同指词语的译文统一、逻辑关系表达等。但是，有一个问题却尚未引起足够重视，即篇章中翻译单位的确定问题。

关于篇章翻译的单位，史晓东曾提出这样的观点："翻译就是求意义等价。而等价的单位是分层次的，体现在音节（音素）、单词、短语、子句、句子、段落、语篇等不同的单位。好的翻译是各个层次都要对等。"（史晓东 2006）这个观点很正确，但不同语言的单位会有很大不同，单位对等就不好办了。尤其是汉语，它的单词、小句、句子都没有清楚的界线。其实，问题的根本还不在于单位边界不清，更基本的问题是这些单位的概念不清。

从人工翻译实践来看，英汉翻译，一般以句号句（句号、叹号、问号为切分符号）为单位；汉英翻译，由于汉语句号的使用没有一定之规，所以常以段落为单位。但是，对于计算机处理来讲，段落肯定是太大了，英文的句号句也是偏大。统计翻译的根基在于有足够数量

*收稿日期：2015 年 6 月 15 日 定稿日期：2015 年 8 月 10 日

基金项目：国家自然科学基金"基于广义话题的汉语篇章结构研究"（61171129）

的有效的双语对齐样本。如果翻译单位选择过大，跨越了小句的连接处，那么比起小句内部，数据样本既要有多倍的长度还要有多倍数量才能反映对译语言之间的词语联系。目前，长句的翻译质量差就是这个问题的反映。

王经益 2009 年从《新概念英语》中抽取了 40 句带有关系从句的句子，调查了华建和 Google 的机器翻译系统。结果是，整句的译文，华建和 Google 的可接受率分别是 35%和 20%；其中主句和从句的翻译，华建和 Google 可接受率分别达 80%和 65%，都远远高出整句翻译的结果（王经益 2009）。这一实验说明长句翻译的确是严重问题。

我们必须确立一种面向篇章的翻译单位体系，这一体系必须受到语言和认知规律的约束，又要顾及机器翻译系统的把控能力。这个问题对于欧洲主要语言之间的翻译也许并不很重要，对于涉及汉语的翻译则具有根本性的意义。

本文的主要目标就是厘清篇章翻译中的翻译单位，进而说明基于这种单位的翻译策略。限于作者语言知识和翻译实践的局限性，本文的工作是以英语和汉语之间的翻译，特别是从英语到汉语的翻译为背景而展开的。

2. 面向机器翻译的翻译单位及其翻译模型

2.1. 篇章翻译单位

我们把篇章翻译单位设计为两个层次，一是基本单位，二是复合单位。复合单位由基本单位构成。从翻译的需要出发，它们应当满足一系列性质：

（1）复合单位是自然状态文本片断。源语言复合单位由篇章直接切割即可得到；目标语言篇章可以由源语言篇章中各复合单位的译文直接接续而生成，除了个别术语、命名实体的译文统一等少数可控的操作之外，基本不加任何变动。（见图 1，其中 Scomp 表示源语言复合单位，Tcomp 表示目标语言复合单位）

（2）基本单位不太大，应在基本的机器翻译技术（统计加规则）的把控范围内；基本单位不太小，其意义和功能具有独立性，使得每个基本单位都能独立地生成确定的译文，并且方便基本单位合成复合单位。

（3）不同的基本单位可能会有共享成分，这种共享关系不能超出复合单位的范围。

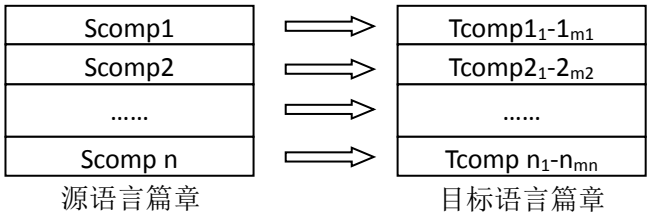


图 1. 在复合单位层面看篇章翻译中原文和译文的关系

2.2. 篇章翻译模型

在这样的单位划分体系之下，篇章机器翻译可以归结为源语言到目标语言的复合单位之间的翻译，其模型可用函数表示：

Ttext=TextTrans(Stext)=TextTrans({Scomp})={CompTrans(Scomp)}　　（公式 1）

其中各符号的含义是：

Ttext：变量，目标语言篇章；Stext：变量，源语言篇章；Scomp：变量，源语言复合单位；TextTrans：函数，篇章翻译，由源语言篇章生成目标语言篇章；CompTrans：函数，复合单位翻译，由源语言复合单位生成目标语言复合单位；{x}：变量 x 的序列。

公式 1 的意思是：

目标语言篇章=源语言篇章的译文=源语言复合单位序列的译文=源语言复合单位译文的序列

其中第一个等号是显然成立的，后两个等号依据上一小节中说的复合单位与篇章的关系。

复合单位的翻译过程可以分成 3 个彼此独立的步骤：

（1）Parsing：源语言文本中复合单位到基本单位序列的拆分；

（2）Translating：源语言基本单位序列到目标语言基本单位序列的翻译；

（3）Assembling：目标语言基本单位序列到复合单位的装配。

这 3 个步骤中，步骤 2 由传统的机器翻译方法实现，步骤 1 和 3 是篇章翻译的特殊要求。

这 3 个步骤可以用下面的复合函数公式来表示：

CompTrans(Scomp)=λ([{Sbasic},Logicrel, Coref])C2BParse(Scomp)

B2CAssemble(BasicTrans({Sbasic}),Logicrel, Coref)　　（公式 2）

其中各符号的含义是：

Sbasic：变量，源语言基本单位；Logicrel：变量，源语言基本单位间的逻辑关系；Coref：变量，源语言基本单位中词语间的同指关系；C2BParse：函数，源语言复合单位的拆分，由 Scomp 生成{Sbasic}、Logicrel、Coref；BasicTrans：函数，基本单位序列的翻译，由{Sbasic}生成目标语言基本单位序列{Tbasic}；B2CAssemble：函数，目标语言复合单位装配，由{Tbasic}、Logicrel、Coref 生成 Tcomp。

式中 λ 是 λ 演算记号，λ([x1,……,Xn])R M 表示一个函数应用。R 是实参；[x1,……,Xn]是形参，表示 R 可以分解为 n 项，各项对应的形参是 x1,……,Xn；M 是函数体。

公式 2 中，C2BParse(Scomp)为实参，其中函数 C2BParse 拆分 Scomp，函数值可分解为 3 部分，分别交给函数体的实参{Sbasic}、Logicrel 和 Coref。函数体中函数 BasicTrans 翻译{Sbasic}，其结果应是目标语言基本单位序列{Tbasic}，它连同 Logicrel 和 Coref 作为函数 B2CAssemble 的自变量，进行目标语言复合单位的装配。

一般来说，目标语言基本单位装配成复合单位，需要依赖目标语言基本单位之间的逻辑语义关系和词语之间的同指关系。在公式 2 中，这两种关系用的是源语言文本中的对应关系。这样的做法是没有问题的，因为这两种关系是在人的认知中存在的，与语言种类无关。

由于基本单位翻译的独立性，源语言基本单位序列的翻译结果就是基本单位翻译结果的序列，因此由公式 2 可以得到公式 3：

CompTrans(Scomp)=λ([{Sbasic},Logicrel, Coreference])C2BParse(Scomp)

B2CAssemble({OneBasicTrans(Sbasic)},Logicrel, Conference)　（公式 3）

其中 OneBasicTrans 是翻译一个基本单位的函数。

这些函数逐个写出来就是：

[{Sbasic}, Logicrel, Coref]=C2BParse(Scomp)

{Tbasic}=BasicTrans({Sbasic})={OneBasicTrans(Sbasic)}

Tcomp=B2CAssemble({Tbasic}, Logicrel, Coref)

其中 Tcomp 是目标语言复合单位，也是函数式 CompTrans(Scomp)的值。

概括来说，这是分拆（Parsing）、翻译（Translating）、装配（Assembling）三步走的模型，简称为 PTA 模型。图示如下：

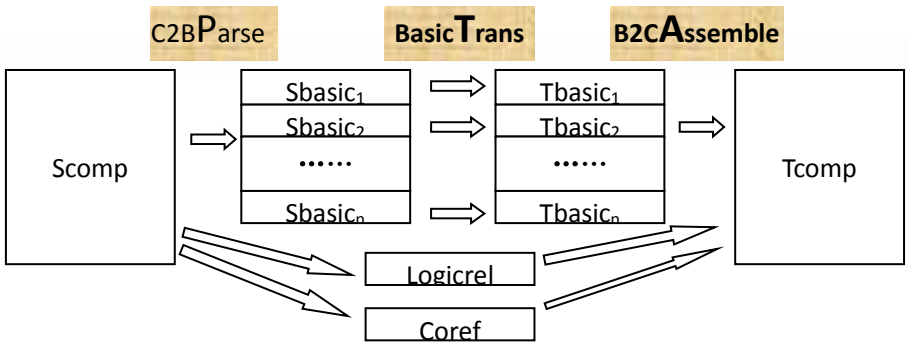


图 2 PTA 模型

上面的函数映射可以用统计方法实现：

Ttext=argmaxP(Ttext | Stext)　　（公式 5）

考虑到复合单位翻译的独立性，有

P(Ttext | Stext)=P({Tcomp} | {Scomp})=∏P(Tcomp | Scomp)　　（公式 6）

由复合单位翻译三个步骤的相互独立性，复合单位翻译的概率是三个条件概率的乘积：

P(Tcomp | Scomp)

=P({Sbasic}, Logicrel, Coref | Scomp)×P({Tbasic}|{Sbasic})×P(Tcomp |{Tbasic}, Logicrel, Coref)

（公式 7）

又由于基本单位翻译的独立性，上述公式中基本单位序列翻译概率又可以进一步分解成各基本单位翻译概率的乘积：

P({Tbasic}|{Sbasic})=∏P(Tbasic | Sbasic)　　（公式 8）

综合公式 5、6、7、8，可以得到篇章翻译的概率计算表达式：

$$Ttext = \arg\max\left(\prod\left(P(\{Sbasic\}, Logicrel, Coref \mid Scomp) \times \prod P(Tbasic \mid Sbasic) \times P(Tcomp \mid \{Tbasic\}, Logicrel, Coref)\right)\right)$$

（公式 9）

其中自变量最大化算子 argmax 是对 Sbasic、Logicrel、Coref、Tbasic 和 Tcomp 而言的，即 3 个概率的计算中要分别对这 5 类变量进行优选。

这里没有涉及源语言篇章 Stext 切分为源语言复合单位序列{Scomp}的步骤。当源语言复合单位有唯一确定的形式标记作为界限时（比如英语以句号为标记），这一步可以忽略不计，否则还需要考虑到模型之中。

基本单位和复合单位的界定是语言相关的。下面分别说明汉语和英语中基本单位和复合单位的界定方法。

3. 汉语翻译单位的界定—广义话题结构和话题自足句

汉语篇章中的标点符号最常见的是逗号和句号，是篇章切分的最重要的形式标记，但不能直接作为翻译单位的切分标记。原因在于逗号句往往信息不全，句号句常常规模太大，而且也可能信息不全。逗号句信息不全和句号句规模太大的实例比比皆是，因篇幅关系这里不予展示，下例是句号句信息不全的实例：

例 1（中华人民共和国宪法第四十一条）

对于公民的申诉、控告或者检举，有关国家机关必须查清事实，负责处理。任何人不得压制和打击报复。

这段话由汉语的两个句号句组成，它的英语译文（引自全国人大网）也是两个句号句：

The State organ concerned must, in a responsible manner and by ascertaining the facts, deal with the complaints, charges or exposures made by citizens. No one may suppress such complaints, charges and exposures or retaliate against the citizens making them.

如果以句号句为单位进行翻译,并且不对第 2 个句号句进行人为的信息补充，不可能译出其中加下划线的部分。

汉语的第 2 个句号句可以看成省略了宾语，但更自然的看法是共享了第 1 个句号句的话题“对于公民的申诉、控告或者检举”。这段话的整体结构可以用以下图式表示：

对于公民的申诉、控告或者检举，

　　有关国家机关必须查清事实，

　　　　负责处理。

　　任何人不得压制和打击报复。

其中包含 3 个单位：

对于公民的申诉、控告或者检举有关国家机关必须查清事实，

对于公民的申诉、控告或者检举有关国家机关必须负责处理。

对于公民的申诉、控告或者检举任何人不得压制和打击报复。

这一图式是汉语篇章的广义话题结构的表示方法。该理论以标点句为单位，以广义话题和说明的语用关系为基本出发点，构建了基于微观话题的汉语篇章静态形式模型；以该形式模型为基础，可以利用上下文补充完整标点句的话题和说明，得到话题自足句，这个操作过程归纳为动态的流水模型。广义话题结构流水模型的核心是堆栈模型，此外有话题后置模型、节栈模型、汇流模型和封闭语段模型（宋柔 2013）。例 1 是堆栈模型的例子，所列的 3 个单位是 3 个话题自足句。

我们以广义话题结构的文本块作为汉语篇章翻译的复合单位，以话题自足句作为汉语篇章翻译的基本单位。它们满足 2.1 节讨论的性质。限于篇幅，这里不予详细讨论。

4. 英语翻译单位的界定—句号句和 NT 小句

我们把英语复合单位界定为句号句，即篇章中由句号、叹号、问号为边界的文本片断。这样的界定显然满足本文 2.1 节所列复合单位的性质。从人工翻译实践看，英语篇章到汉语篇章的技术性的人工翻译（区别于文学性的人工翻译），通常的确是以句号句为篇章的下一级单位。

英语篇章基本单位的确定是比较复杂的。

例 2（华尔街日报）

Documents filed with the Securities and Exchange Commission on the pending spinoff disclosed that Cray Research Inc. will withdraw the almost $ 100 million in financing it is providing the new firm if Mr. Cray leaves or if the product-design project he heads is scrapped .

百度译文：

提交给美国证券交易委员会对未决分拆披露，克雷研究公司将收回近 100000000 美元的融资，这是为新公司如果克雷离开或者产品设计项目他头报废文件。

百度翻译没做好的地方有两方面：

（1）未能识别名词短语与其后修饰语的关系，造成译文混乱。涉及这类错误的成分是：

- 名词“Documents”的过去分词后修饰语“filed with the Securities and Exchange Commission”；
- 名词“Documents”的介词短语后修饰语“on the pending spinoff”；
- 名词短语“the almost $ 100 million in financing”的关系从句“it is providing the new firm”
- 名词短语“the product-design project”的关系从句“he heads”

（2）状语从句译文顺序不对。涉及这类错误的成分是：

- 主句“Cray Research Inc. will withdraw the almost $ 100 million in financing”的状语从句“if Mr. Cray leaves”
- 同一个主句的另一个并列的状语从句“if the product-design project is scrapped”

这些错误给我们两条启发：

（1）名词短语和它的后修饰成分应当从句号句中提取出来，二者结合作为翻译的基本单位。需要说明这种结合的关系类型，以便翻译并做译文装配。

关于这一类基本单位，更准确的说法是指称语加它的后修饰成分，因为作为被修饰对象的不全是名词短语，有可能是形容词短语、非限定的动词短语或者主谓结构的小句，它们在受到后修饰时，实际的语用功能不是陈述或修饰，而是指称。

（2）状语从句应当从句号句中提取出来作为翻译的基本单位。需要说明是哪个主句的从句，以便译文在装配时调序。

举一反三，可以想到：

（3）宾语从句应当成为一个基本单位。需要说明主从关系，以便译文在装配时安排顺序。

（4）并列的小句应当分别作为一个基本单位。需要说明与哪个几个小句并列，以便译文在装配时安排顺序。。

（5）上述各类提取出来准备作为翻译基本单位的成分，如其中又含有这些类成分中的某一类或某几类，也应提取出来作为翻译基本单位。因此这一过程是递归的，直至没有这些类成分需要提取，而最后剩下的也是翻译基本单位，它应当是简单的主谓关系小句。

以上就是英语篇章翻译的基本单位的分类体系。其中第（1）类还要加以细化，因为有相当多的名词后修饰成分只有一两个词，如 N of N 等，它们通常在传统的基于句子的机器翻译系统的掌控范围内，而且单独翻译有歧义，并致译文装配繁琐，因此不应单独取出来做基本单位。下面是我们迄今为止在语料库标注中归纳出的细化类型。

（1.1）指称语++关系从句

例 3：The missing watch is emblematic of the problems Mr. Wathen encountered .

其中 the problems++Mr. Wathen encountered 是一个基本单位，类型为 WO，表示该基本单位是关系从句类型，先行语在关系从句中充任宾语。原句中去掉这个基本单位后剩下的成分 The missing watch is emblematic of the problems 也是一个基本单位，类型为 SV，即为简单的主谓结构小句。注意，the problems 是这两个基本单位的共享成分。

指称语++关系从句是一个大类，包括几个小类。除了 WO 类外，还有 WS 类，表示先行语在关系从句中充任主语；WC 类表示先行语在关系从句中充任主语，而且先行语本身是一个小句；WD 类表示先行语在关系从句中充任状语；WPO 类表示先行语在关系从句中充任介词宾语；WE 类表示先行语和从句等同，形式主语 it 引导的从句属于这一类。

本节的例子都来自华尔街日报宾州树库。下面的例子不再指明去掉所例示的基本单位后剩下的成分，也不再指明共享成分。

（1.2）指称语++过去分词短语

例 4：And though the size of the loan guarantees approved yesterday is significant.

其中 the loan guarantees++ approved yesterday 是一个基本单位，类型为 ED，表示指称语的后修饰成分是过去分词短语。

（1.3）指称语++现在分词短语

例 5：Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer .

其中 Four of the five surviving workers++ including three with recently diagnosed cancer 是一个基本单位，类型为 ING，表示指称语的后修饰成分是现在分词短语。

（1.4）指称语++动词不定式短语

例 6：The plant , which is owned by Hollingsworth &Vose Co. , was under contract with Lorillard to make the cigarette filters .

其中 Lorillard++ to make the cigarette filters 是一个基本单位，类型为 TO，表示指称语的后修饰成分是动词不定式短语。

（1.5）指称语+形容词短语

例 7：They had all maintained with a certain fidelity a manner of technique and composition consistent with those of America’s first popular landscape artist .

其中 technique and composition++consistent with those of America’s first popular landscape artist 是一个基本单位，类型为 ADJ，表示指称语的后修饰成分是形容词短语。

（1.6）指称语++介词短语

例 8：The survival of spinoff Cray Computer SEQp. as a fledgling in the supercomputer business appears to depend heavily on the creativity .

其中 spinoff Cray Computer SEQp. ++ as a fledgling in the supercomputer business 是一个基

本单位，类型为 PPM，表示指称语的后修饰成分是嵌套的介词短语，即介词的宾语又带有后修饰语。

指称语++介词短语是一个大类，包括几个小类。除了 PPM 类外，还有 PPI 类，表示动名词短语作介词宾语；PPP 类，表示过去分词短语作介词宾语。名词短语的后修饰成分是介词短语的情况很多，我们只把这 3 类归入基本单位，一是因为这 3 类的介词短语往往比较长，如不从句子中取出，会使整个句子太长，以致传统的面向句子的机器翻译系统无法把控；二是因为作为中心语的后修饰成分这 3 类介词短语，有比较强的陈述性，能被独立翻译。

（1.7）指称语++同位语

例 9：Rep. Jerry Lewis , a conservative Californian , added a provision of his own intended to assist Bolivia ,

其中 Rep. Jerry Lewis++a conservative Californian 是一个基本单位，类型为 APP，表示指称语的后修饰成分是指称语的同位语，用以对指称语进行解释。

（1.8）指称语++插入的后修饰成分

例 10：But maintaining the key components of his strategy -- a stable exchange rate and high level of imports -- will consume enormous amounts of foreign exchange .

其中 the key components of his strategy++a stable exchange rate and high level of imports 是一个基本单位，类型为 EXP，表示指称语的后修饰成分是一个插入语，对指称语做进一步的解释。

上面是第（1）大类的细分，它们都具有指称语+后修饰成分的结构。这些后修饰成分，无论是关系从句、动词短语、形容词短语还是插入语，以及以动词短语为宾语的介词短语，都具有明显的陈述意义。嵌套的介词短语以及同位语虽然通常是静态的修饰，但也可看成是对于前面指称语的属性描述，从而也具有陈述意义。比如例 9 中同位语类型的 Rep. Jerry Lewis++a conservative Californian，意思是“众议员 Jerry Lewis 是加州的保守党人”。因此，这一大类的后修饰成分都可以看成陈述语。（1）至（5）各类样例取出尽可能多的基本单位最后剩下的成分，都是主谓结构的小句，主语是指称语，谓语也是陈述语。所以，英语篇章翻译的所有基本单位，都具有指称语+陈述语的结构。但这里说的指称语和陈述语的概念与语言学中的传统概念有所不同。为了在加以区分同时还表现出这些结构的特点，我们借用儿童语言习得中的术语，把这里的指称语称为 naming part，陈述语称为 telling part，一个基本单位就是一个 naming-telling 结构的小句，简称为 NT 小句。

5. NT 小句的性质

如第 3 节所述，汉语篇章翻译的基本单位是广义话题结构的话题自足句。每个话题自足句都具有话题-说明结构，话题和说明可以分别看成 naming part 和 telling part，因此汉语的话题自足句其实也是 NT 小句，汉语广义话题结构文本可以拆分成 NT 小句的序列。因此，NT 小句是英语和汉语篇章翻译的同类型的基本单位。

据涉及多种语体的数万汉语 NT 小句和数千英语 NT 小句的调查，NT 小句具有如下性质：

（1）意义和功能的独立性和完整性。NT 小句能相对于上下文基本上独立完整地表达意义。独立性指的是它的意义不受上下文的影响而改变。完整性指的是它能表达相对完整的事理意义，包括事物的性状、事件或关系的基本要素。NT 小句装配成复合单位，只需要遵从逻辑顺序，无需再考虑事理意义所致的句法关系。

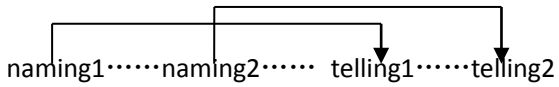
（2）全覆盖性。任何正常篇章中的所有成分，除了一些明显的插入性成分和句首的一些状性成分之外，基本上都可划入某一个 NT 小句的 naming part 或 telling part。

（3）顺序性。除了话题后置的情况，naming part 总是在 telling part 前面。而话题后置有比较严格的约束条件。

（4）分支性。除了极特殊的情况（非主谓句和主语省略句），每个 naming part 有至少一个

直接的 telling part，每个 telling part 恰有一个直接的 naming part。因此，会有多个 NT 小句共享一个 naming part 的情况。

（5）不可交叉性。除了极特殊的情况（具有插入附注性质的背景语段和补充说明语段），各 NT 小句之间互不交叉穿越。即不存在具有如下形式的文本片断：



（6）成句性。NT 小句基本成句。每一个这种结构或者本身就是通常意义下的句子，或者简单地插入删除一两个用于连接的成分就可以变成通常意义下的句子。

（7）非递归性。每一个 NT 小句的内部构造基本非递归，长度受限。

上面性质中，第 1 条满足了作为篇章分析基本单位的基本要求；第 2 条是这种结构的应用基础；第 3、4、5 条保证了 NT 小句分析的可操作性；第 6 条和第 7 条是小句快速认知的基础，也是小句自动翻译的基础。

人对客观世界认知的出发点就是 naming part，认知的内容就是 telling part，因此 NT 小句是人类认知的基本单位。

以上一系列分析说明，NT 小句适合于表示篇章结构，能反映认知约束，具有高度的实证性，因此具有篇章文法的意义。NT 小句满足篇章翻译基本单位的要求，能有效支持面向篇章的机器翻译。

6. 英汉篇章翻译实例

我们以例 2 为例，展示在这样的翻译单位体系下，采用三步走的 PTA 模型，如何完成英汉篇章翻译工作。

英语原句

Documents filed with the Securities and Exchange Commission on the pending spinoffdisclosed that Cray Research Inc.will withdraw the almost $ 100 million in financingit is providing the new firm if Mr. Cray leaves or if the product-design projecthe headsis scrapped .

第一步 Parsing：英语句号句拆分为 NT 小句

（1）英语名词为中心的结构分析

首先采用换行缩进图式表示英语的 naming part 和 telling part。每个 telling part 换行缩进到它对应的 naming part 右下方。为了标示 naming part 的左边界，需要在 naming part 的左下方加注双竖线，但当 naming part 在行的左端时就无需双竖线标记了。

```
Documents//NAMING
        filed with the Securities and Exchange Commission //ED
        on the pending spinoff //PP
        disclosed that//SV-OBJ
【Cray Research Inc.will withdraw    the almost $ 100 million in financing //SV
                                ‖                                    it is providing the new firm //WO
if Mr. Cray leaves//SV
or if    the product-design project //SV
   ‖                          he heads //WO
                              is scrapped .//SV】
```

注：

1) 每个 telling part 的右端标注它所对应的 NT 小句类型。

2) 第 3 行只是一个非嵌套的介词短语，本来不必列做一个 telling part，但是第 2 行的成分列做 telling part 单独成行后，这个介词短语就被孤立出来了，所以只能单独成行，与它修饰的 documents 构成一个 NT 小句，标注为 PP 类型，表示是介词短语。

3) 第 4 行对应的 NT 小句是一个主谓结构，但谓语动词缺少宾语从句，所以类型标注为 SV-OBJ。

4) 图式中的黑方括号括起分列多行的宾语从句。

NT 小句的不可交叉性保证了英语的句号句能够用这种换行缩进图式表示，进而能由此图式机械地生成 NT 小句。

（2）英语小句拆分、逻辑分析和同指分析

这一小步将上一步分析出来的 naming part 及对应的 telling part 放到一起，构成 NT 小句，并表示出小句间的成分共享关系、逻辑关系、同指关系。

⑴(Documents)1 +ED+ filed with the Securities and Exchange Commission

⑵(Documents)1 +PP+ on the pending spinoff

⑶(Documents)1 disclosed that

⑷【{Cray Research Inc.}4+SV+will withdraw (the almost $ 100 million in financing)2

⑸ (the almost $ 100 million in financing)2 +WO+ {it}4 is providing the new firm

⑹ if {Mr. Cray}5+SV+ leaves

⑺ (the product-design project)3 +WO+{he}5 heads

⑻ or if (the product-design project)3+SV+is scrapped .】

(OBJ (ATT（⑴）(⑵）(⑶)) (SUB (PAR（⑹）(ATT(⑺）(⑻)) (ATT（⑸）(⑷)))

注：

1) 各 NT 小句的类型标注在连接 naming part 和 telling part 的两个加号中间。

2) 有些 naming part 被多个 telling part 共享，于是需要复制多份。为了表明被复制成分的同一性，这些成分用圆括号括起来并加标数字。

3) 在英语原文中，有些概念初次出现时用名词表示，再次出现时用代词表示。但是，英语 NT 小句的译文在组成汉语广义话题结构时，有时需要调序。这种调序有可能把同一概念的代词性出现调到了名词性出现的前面，但这是不合乎汉语篇章语法的。为此，在英语处理阶段，需要把代词的同指对象识别并标注出来，以便在汉语装配阶段配合小句调序进行指代方式的调整。标注方法是花括号右边加数字，标示指代的同一性。

4) 逻辑分析的目的是装配阶段能将各小句译文按照汉语的逻辑顺序排列好。上面最后一行是小句间的逻辑关系式，格式为（逻辑关系符 {逻辑前项行号}+ 逻辑后项行号）。本例中使用的逻辑关系符有：OBJ-主句和宾语从句，ATT-修饰成分和被修饰成分所在句，SUB-主从复句中的从句和主句，PAR-并列复句，IND-独立小句。

第二步 Translating：英语 NT 小句到汉语 NT 小句的翻译

（3）英汉小句翻译（同一原文的译文应当相同）

⑴(文件)1+ED+提交给了证券交易委员会

⑵(文件)1+PP+关于这个未决分拆

⑶(文件)1 透露

⑷【{克雷研究公司}4+SV+将撤回(将近 1 亿美元的融资)2

⑸ (将近 1 亿美元的融资)2+WO+{它}4 正提供给该新公司

⑹ 如果{克雷先生}5+SV+离开

⑺ (产品设计项目)3+WO+{他}5 带领

⑻ 或者如果(产品设计项目)3+SV+被废止】

(OBJ (ATT (1) (2) (3)) (SUB (PAR (6) (ATT(7) (8))) (ATT (5) (4)))

第三步 Assembling：汉语 NT 小句装配成汉语广义话题结构的文本

（4）汉语说明语处理

⑴[关于这个未决分拆⑵]<的> (文件)1 提交给了证券交易委员会

⑶ (文件)1 透露

⑷【{克雷研究公司}4 将撤回(将近 1 亿美元的融资)2

⑸ {它}4 正把(将近 1 亿美元的融资)2 提供给该新公司

⑹ 如果{克雷先生}5 离开

⑻ 或者如果[{他}5 带领⑺]<的> (产品设计项目)3 被废止】

(IND (1) (OBJ (3) (IND (5) (SUB (PAR (6) (8)) (4))))

注：这一小步变化如下：

1) 长的 NT 小句⑴和⑸改造成普通句，进而句⑵的 telling part 嵌入被说明成分所在小句⑴中，插在被说明成分前，标上原来小句的序号⑵，中间加“的”。

2) 表示修饰关系的 NT 小句⑺中的 telling part 嵌入到被修饰成分所在小句（8）中，插在被修饰成分前，标上原来小句的序号⑺，中间加 “的”。

3) 小句间逻辑语义关系需要修正，具体过程不详述。

（5）汉语小句调序和指代变换

⑴[关于这个未决分拆⑵]<的> (文件)1 提交给了证券交易委员会

⑶(文件)1 透露

⑸【{克雷研究公司}4 正把(将近 1 亿美元的融资)2 提供给该新公司

⑹ 如果{克雷先生}5 离开

⑻ 或者如果[{他}5 带领⑺]<的>(产品设计项目)3 被废止

⑷ {它}4 将撤回(将近 1 亿美元的融资)2。】

(IND (1) (OBJ (3) (IND (5) (SUB (PAR (6) (8)) (4))))

注：这一小步中的变化如下：

1) 按照逻辑关系式中小句序号的线性顺序，对小句序列进行调序，第⑷小句移到最后。

2) 因⑷移到后面，故⑷和⑸中同指的名词形式和代词形式对调。

（6）汉语小句删除共享成分，加标点并删除标记得到最后结果

关于这个未决分拆的文件提交给了证券交易委员会。

该文件透露：

【克雷研究公司正把将近 1 亿美元的融资提供给该新公司。

如果克雷先生离开，

或者他带领的产品设计项目被废止，

它将撤回这笔融资。】

注：1）第 5 行与第 4 行中的“如果”，可以作为广义话题共享而删除。

2）独立句后为句号，带宾语从句的主句后为冒号，并列复句中间用逗号，主从复句的从句和主句之间用逗号，复句结束用句号。

7. 英汉篇章 NT 小句对齐语料库建设

我们正在进行英汉篇章 NT 小句对齐语料库的建设工作。源语料是宾州英语树库中华尔街日报（wsj）树库的英语原文，标注内容是按 PTA 模型的三大步展示每个英语句号句翻译成汉语广义话题结构文本的翻译过程。具体来说，有精加工和粗加工两个方案。

精加工方案：

第一步加工英语篇章。以句号句为单位，采用换行缩进图式的直观方式，展示 naming part 和 telling part 的关系，构造 NT 小句，标注 NT 小句的类型、NT 小句之间的逻辑关系、复制成分的同一性关系、代词的同指关系。

第二步进行英语 NT 小句到汉语 NT 小句的翻译。这步翻译采用机助人译的方式，在字面忠实于原文的前提下要求通顺，以适应机器学习的需要。

第三步进行汉语 NT 小句到汉语篇章广义话题结构的装配，装配操作包括：某些 NT 小句变成普通句(话题自足句)，某些 NT 小句的 telling part 嵌入到另一个 NT 小句中 naming part 之前，按小句间的逻辑关系进行小句调序，按小句调序结果修改指代方式，广义话题结构中的共享成分删除，按逻辑关系加标点。

粗加工方案：

第一步加工英语篇章。以句号句为单位，采用换行缩进图式的直观方式，展示 naming part 和 telling part 的关系，对于每个 telling part 标注它所对应的 NT 小句类型。

第二步翻译。以机助人译的方式在换行缩进图式中将英语的每个 naming part 和 telling part 翻译成汉语。

第三步汉语装配。人工将换行缩进图式中的汉语片断装配成汉语文本。

目前，粗加工的英汉小句对齐语料已有数千句号句，为语料的精加工奠定了一定的基础，下一步精加工结果可以由粗加工结果扩展而得到。

语料库将显示粗加工和精加工的每一大步的结果，以及精加工的每一小步的结果，以便机器学习和人工总结规律。

英汉篇章 NT 小句对齐语料库的建设目的有以下几方面：

- 直接用作英汉机器翻译的训练语料。
- 完善机器翻译的评价系统。
- 面向人的英汉翻译研究。
- 在 NT 小句体系下，跨语言的篇章语法研究，英语和汉语的小句异同研究
- 语言认知机制和认知语法研究。
- 语言习得和语言教学研究。

8. 关于 PTA 模型的讨论

8.1. PTA 模型的来由

PTA 模型并非无本之木。王力先生 70 多年前在西南联大《中国现代语法》的讲义中专门讲了欧化的语法（王力，1954）。他举了多个英汉长句翻译的实例，其中一个是

People who have enjoyed good educational opportunities ought to show it in their conduct and language.

他说这句话的欧化译法是“已经享受过良好教育机会的人们应该在他们的行为言语上表现它。”，非欧化的译法是“一个人享受过良好教育的机会，应该在行为和言语上表现出来。”从王力的处理方法可以看出，他把英语的主句和关系从句拆分开来，分别翻译，然后使用两种装配方法，一是用从句译文作为先行语的前置修饰，一是将从句译文另置一句，并共享主句话题。就这个例子来说，后一种译法更自然。

其实，一般人在英汉人工翻译的实践中都有这样的思路，即将主句和从句以及一些附带的修饰成分分开翻译，再看如何装配更符合汉语习惯。PTA 模型是人的认知的抽象。

基于规则的机器翻译，效果较好的方法是转换翻译（Hutchins2009），一般划分为三个阶段：原文分析、原文译文转换和译文生成（冯志伟 2012）。这个模式与本文的 PTA 模型思想相近，只是转换翻译是面向句子的，分析的叶结点是词；PTA 模型是面向篇章的，分析的叶节点是 NT 小句。这是机器翻译方法的螺旋式的上升。

8.2. PTA 模型的可行性分析

PTA 模型的三个步骤--拆分、翻译和装配，每一步都有很大的难度。假如说每一步的准确率都达到 70%，由于这 3 步相互独立，最终的翻译准确率只有 70%的 3 次方，不到 35%。如此看来，这个方法似乎是没有前途的。

实际情况是，任何一个复杂问题的求解，都应尽可能地分解为多个子任务，而且最好子任务之间相互独立。PTA 模型的 3 个步骤正是对篇章翻译这一复杂任务的分解，而且分解出的这 3 个子任务确实相互独立。这 3 个子任务各自面对的困难都是篇章翻译原来就有的，只不过目前的一般方法是混在一起处理，PTA 模型则是把困难分开，分别处理。混在一起处理，某些实例可能处理得好，某些实例可能处理得不好，这些处理结果都带有一定的偶然性，难以把控，从而也就难以提高。分别处理的方法可以减少偶然性，可以针对不同困难的特点各个击破。

PTA 方法的基础是人对于长句翻译方法的理性认知，即拆成 NT 小句分别翻译后再装配。这种宏观层面的理性认知不同于细节处理的规则，它具有高度的概括性和有效性。至于 3 个步骤中的每一步，既可以使用理性的知识，也可以使用来自语料库数据的经验，还可以把二者结合起来。如果采用机器学习的方法，由于每个子任务的都比篇章翻译的总任务来得简单，目标较为集中，需要的特征比较少，机器学习的效果就会比较好。

进一步分析这一方法的难度：

- 拆分和装配只涉及单一的语言，可以在巨大的单语语料库中学习。
- 翻译涉及双语，但学习对象是 NT 小句，大致是非递归的简单句，长度短，语法语义模式的重复率会比较高，而且短句对齐会比较容易，如此自然有利于机器学习。
- 拆分中要做 NT 小句识别，这基本上是一种浅层分析，难度不会太大。
- NT 小句间逻辑关系分析和词语同指关系分析为的是汉语装配中的调序。目标清楚，所以任务得以限定：逻辑关系分析中主从关系不必分出细类，因为汉语的句序在无连词的请况下通常是从句在主句前，与主从关系的细类无关；同指分析仅限于有主从关系而且主在后从在前的小句，因为其他小句不涉及调序，无需作指代变换。退一万步讲，可以不做逻辑关系分析和同指分析。只要 NT 小句识别正确，将小句译文直接连起来作为装配结果，也能基本明白意思，总比各种小句混在一起乱成一锅粥要强。

本文作者曾指导博士生王经益研究英语带有关系从句的复杂长句遵从 PTA 模型的汉译，主要做译文的装配。王经益归纳出了一系列特征和规则，在《新概念英语》的开放测试中，对于从句译文应当前置修饰还是另置一句，区分的正确率接近 80%（王经益 2009）。

目前，作者的课题组正依照 PTA 模型进行的英汉小句对齐语料的建设。当前的首要目标是通过人工标注，完善这一方法的理论基础和形式模型。虽然机器翻译的实验尚未进行，但数千句号句英汉翻译的人工考察，支持了该模型的可行性。

本文的目标首先是英汉翻译。理论上说，汉英翻译也可以照此模式进行。我们目前暂未实施，原因在于这三步的技术成熟程度还不够。

第一步，汉语缺少形式标记，汉语广义话题结构的分析较之英语 NT 小句分析更为困难；

第二步，汉语小句翻成英语小句，可用于机器学习的高质量的对齐语料较英到汉更为缺乏，而且从形式标记少的语言翻译成形式标记多的语言本身难度就大于反向翻译。

第三步，英语小句的排序和改造，对于英语语感的要求非常高，通常需要具备较高语言

学修养的英语母语者来完成，而我们的语料库的建设人员通常只是汉语母语者。由于这一步的结果将用作机器翻译的训练样本，要求质量高，故必须慎重。

用 PTA 模型做汉英翻译的工作，是一个合理的目标，也是我们后期的工作方向。为此，必须在这三步上多加努力，尤其要加紧汉语广义话题结构分析的研究工作。

参考文献

[1] 冯志伟. 《统计机器翻译》述评[J]. 外语教学与研究，2003，45（4）：629-633.

[2] 冯志伟. 《统计机器翻译》序[M]. 北京：电子工业出版社，2012：3-14.

[3] 刘群. 机器翻译技术现状与展望[J]. 集成技术，2012，1（1）：48-54.

[4] 宗成庆. 统计自然语言处理（第二版）[M]. 北京：清华大学出版社，2013.

[5] 张民. 语义、语篇和机器翻译[R]. 贵阳：CIPSC 战略研讨会，2014.

[6] 史晓东，陈毅东. 基于语篇的机器翻译前瞻[C]. 曹右琦，孙茂松. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议，北京：清华大学出版社，2006：34

[7] 宋柔. 汉语篇章广义话题结构的流水模型[J]. 中国语文，2013（6）：483-494.

[8] 王力. 中国语法理论[M]. 北京：中华书局. 2012：352-354

[9] Hutchins, J. Machine translation General overview [A]. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics* [C]. Beijing: Foreign Language and Research Press & Oxford University Press, 2009.

[10] 王经益. 面向计算机的英语关系从句汉译研究[D]. 北京：北京语言大学，2009